REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
MARCH 2009	Conference Paper Preprint	March 2008 – April 2009

4. TITLE AND SUBTITLE

DIALECT DISTANCE ASSESSMENT BASED ON 2-DIMENSIONAL PITCH SLOPE FEATURES AND KULLBACK LEIBLER DIVERGENCES (PREPRINT)

5a. CONTRACT NUMBER
FA8750-05-C-0029 & 09-C-0067

5b. GRANT NUMBER
N/A

5c. PROGRAM ELEMENT NUMBER
35885G

6. AUTHOR(S)

Mahnoosh Mehrabani, Hynek Bořil, and John H.L. Hansen

5d. PROJECT NUMBER
3188

5e. TASK NUMBER
BA

5f. WORK UNIT NUMBER
AE

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Research Associates for Defense Conversion, Inc.
10002 Hillside Terrace
Marcy, NY 13403-2102

CRSS
University of Texas at Dallas
Richardson, TX 75083

8. PERFORMING ORGANIZATION REPORT NUMBER
N/A

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

AFRL/RIEC
525 Brooks Road
Rome NY 13441-4505

10. SPONSOR/MONITOR'S ACRONYM(S)
N/A

11. SPONSORING/MONITORING AGENCY REPORT NUMBER
AFRL-RI-RS-TP-2010-7

12. DISTRIBUTION AVAILABILITY STATEMENT

Approved for public release; distribution unlimited. PA Case # 88ABW-2009-1420, Date Cleared: 08-April-2009

13. SUPPLEMENTARY NOTES

This work, resulting in whole or in part from Dept of Air Force contract number(s) 05-C-0029 & 09-C-0067, has been submitted and accepted for publication in the proceedings of the ICASSP 2010, International Conference on Acoustics, Speech, and Signal Processing, March 2010. If published, publisher may assert copyright. The United States has for itself and others acting on its behalf an unlimited, paid-up, nonexclusive, irrevocable worldwide license to use, modify, reproduce, release, perform, display, or disclose the work by or on behalf of the Government. All other rights are reserved by the copyright owner.

14. ABSTRACT

Dialect variations of a language have a severe impact on the performance of speech systems. Knowing how close or separate dialects are in a given language space provides useful information to predict or improve, system performance when there is a mismatch between train and test data. Distance measures have been used in several applications of speech processing, including speech recognition, speech coding, and speech synthesis. Apart from phonetic measures, little if any work has been done on dialect distance measurement. This method of dialect separation assessment based on modeling 2D pitch slope patterns within dialects is proposed. Kullback-Leibler divergence is employed to compare the obtained statistical models. The presented scheme is evaluated on a corpus of Arabic dialects. The sensitivity of the proposed measure to changes on input data is quantified. It is also shown in a perceptive evaluation that the presented objective approach of dialect distance measurement correlates well with subjective distances.

15. SUBJECT TERMS

Distance measure, separation assessment, dialect classification, pitch features

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			John G. Parker, Jr.
U	U	U	UU	5	19b. TELEPHONE NUMBER (*Include area code*) N/A

Dialect Distance Assessment Based on 2-Dimensional Pitch Slope Features and Kullback Leibler Divergence

*Mahnoosh Mehrabani, Hynek Bořil, John H.L. Hansen**

Center for Robust Speech Systems (CRSS)
Erik Jonsson School of Engineering & Computer Science, University of Texas at Dallas, U.S.A
mahmehrabani@student.utdallas.edu, hynek@utdallas.edu, john.hansen@utdallas.edu

Abstract

Dialect variations of a language have a severe impact on the performance of speech systems. Therefore, knowing how close or separate dialects are in a given language space provides useful information to predict, or improve, system performance when there is a mismatch between train and test data. Distance measures have been used in several applications of speech processing, including speech recognition, speech coding, and speech synthesis. However, apart from phonetic measures, little if any work has been done on dialect distance measurement. This study explores pitch movement differences among dialects. A method of dialect separation assessment based on modeling 2D pitch slope patterns within dialects is proposed. Kullback-Leibler divergence is employed to compare the obtained statistical models.The presented scheme is evaluated on a corpus of Arabic dialects. The sensitivity of the proposed measure to changes on input data is quantified. It is also shown in a perceptive evaluation that the presented objective approach of dialect distance measurement correlates well with subjective distances.
Index Terms: distance measure, separation assessment, dialect classification, pitch features

1. Introduction

Dialect is a variety of a language that is used by a group of speakers belonging to some geographical region. Dialects of a language differ in phonetic, grammatical, and lexical features. The distinction between a dialect and a language is sometimes contradictory. Mutual comprehensibility is a primary criterion for distinguishing a dialect from a language. Unlike speakers of different languages, speakers of different dialects of a language generally understand each other, even with some difficulty [1]. Like any other speaker variation, dialect impacts the performance of speech systems. Therefore, efficient dialect classification algorithms will contribute to improved speech recognition, speaker identification, speech coding, or spoken document retrieval systems. Compared to language identification in which a dictionary and set of language rules are known, dialect classification is more challenging. In a dialect ID task, dialect-dependent models are trained , and during the test phase, the model which is most likely to produce the test utterance is identified. For both train and test phases, feature vectors are extracted from audio files. The availability of data transcription influences the design of dialect ID system. For unsupervised dialect classification, systems based on Gaussian Mixture Models (GMMs) have proven to be successful [2, 3].

*This project was funded by AFRL under a subcontract to RADC Inc. under FA8750-05-C-0029, and the University of Texas at Dallas under Project EMMITT.

Distance measures have been applied in different fields of speech processing. In speech recognition, from measuring the distortion between input and output [4, 5, 6] to speaker adaptation and speaker clustering [7, 8], measures of similarity have played a significant role in improving the system's performance. Other areas of speech processing, such as speech coding, enhancement, and synthesis have exploited distance as objective measure of assessing speech quality [9, 10]. Similarities between different languages have also been utilized for multilingual phoneme modeling [11].

In this study, our focus is on estimating the separation between different dialects of the same language. Phonetic distance between dialects have been calculated in several linguistic studies using various string distances including Levenshtein, Euclidean, and Manhattan distance [12]. The obtained distances have been applied in order to divide geographical maps into dialect areas. Apart from the linguistic approaches, little if any work has been done on finding a distance measure between dialects. In this paper we propose a probabilistic method to compare dialect models trained by 2-dimentional pitch slope features. The advantage of proposed method is that it compares pitch patterns in different dialects, using conversational speech with no transcription in an unsupervised manner. The proposed dialect distance assessment framework, which is based on the available train data, shows how accurately the dialects can be distinguished. Therefore, it provides some sense of the resulting dialect classification system performance, as well as taking an initial step towards dialect purity assessment. Furthermore, the performance of a dialect-dependent speech recognition system for a new dialect can be estimated based on the distance between dialects. In a previous study [13], we assessed dialect separation comparing log-likelihood score distributions. GMMs were applied as statistical models for each dialect, and Mel Frequency Cepstral Coefficients (MFCCs) were used as extracted features from audio files. In this study we show that the pitch pattern based separation assessment is consistent with the log-likelihood score distribution distance for the same corpus. We present the proposed distance measure in Sec. 2. In Sec. 3 the results and their evaluation on a corpus of Arabic dialects are discussed. We also show the repeatability of presented measure, and its correlation with human perception. Conclusions are drawn in Sec. 4. For the remainder of this paper we use "dialect distance" and "dialect separation" interchangeably, and the word "distance" is not used in the strict sense of metric spaces.

2. Proposed Method

Human perception tests indicate that prosodic cues, including pitch movements, can be employed to distinguish one language or accent from another [14, 15]. However, prosodic features have not been fully exploited in language ID systems [16], as well as in dialect classification. The trm "pitch" represents the perceived fundamental frequency of voiced speech. The pitch variation while speaking (intonation) is an aspect of speech which varies among dialects. Therefore, changes in pitch can be applied to dialect separation assessment.

2.1. 2-Dimensional Pitch Slopes

As we briefly mentioned in the introduction section, our objective is to develop an unsupervised system that automatically assesses the separation between dialects, based on available train data. The system's input is untranscribed conversational audio, and we want to compare different dialects on the basis of pitch movements. Therefore, our approach is to statistically model pitch changes in voiced speech data for each dialect, with a fixed pitch feature vector length. Note that our efforts focus on unrestricted data which represents unknown speakers, and unknown text. As the first step, pitch frequencies are extracted from every utterance of each dialect using Robust Algorithm for Pitch Tracking (RAPT) [17] to obtain a single pitch vector per utterance. Next, pitch vectors are normalized on the utterance level by subtracting the utterance's mean pitch in order to reduce inter-speaker pitch variability. 3-Dimensional feature vectors are then generated from all the three consecutive nonzero pitch values. Since we are looking for features that show the changes in pitch, instead of raw pitch we use the slope. The step size in pitch extraction algorithm is fixed (10 msec.), therefore every two pitches are subtracted to obtain pitch slope or delta pitch. Consequently, from each 3D feature vector [F_{0_1} F_{0_2} F_{0_3}], 2D vector [(F_{0_2}-F_{0_1}) (F_{0_3}-F_{0_2})] is derived.

In the next step, all 2D pitch slope feature vectors extracted from every speaker and utterance of each dialect's data are used to build 3D histograms. Each histogram is considered as a statistical representation of pitch change in the corresponding dialect. Fig. 1 shows an example of 3D pitch slope histogram. Each 2D pitch slope vector corresponds to a point on the XY plane.

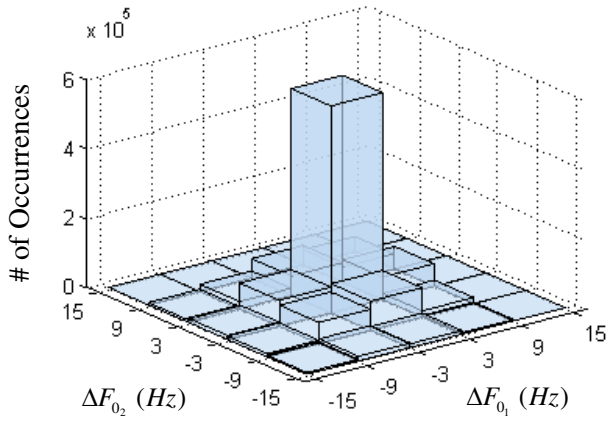

Figure 1: *3D histogram for Egyptian Arabic dialect pitch slopes.*

2.2. Pitch Pattern Codebook

Now that we have extracted all the 2D pitch slopes for each dialect, the next step is to find the pattern of changes in every three consecutive pitches. For instance, a positive slope means an increase in pitch, and alternatively a negative slope represents a decrease. However, the absolute value of the slope or pitch change is also important. In other words, steep slopes correspond to abrupt changes in pitch. We experimentally set some thresholds for delta pitch to obtain a codebook of pitch patterns. 9 different patterns are considered for each dialect which are depicted in Fig. 2. If the absolute change of pitch is less than 3 Hz, the pitch is considered to be almost unchanged. However, for absolute pitch slopes more than 3, two options are considered: positive and negative. Next, for each pattern, the probability of occurrence in the given dialect is calculated. This way, we build a statistical model for pitch patterns which in fact is a discrete probability distribution. The pitch pattern model for each dialect can be described by matrix $P(3 \times 3)$ of probabilities.

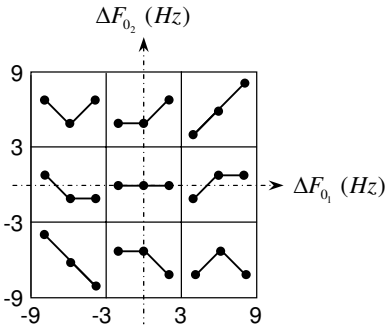

Figure 2: *Pitch slope patterns.*

Furthermore, the obtained pitch patterns of 3D pitch vectors, can be extended to higher dimensions using N-grams. In order to build the codebook of patterns for 4D pitch vectors, bi-grams are computed. We first consider each of the 9 pitch patterns as a word in dictionary: $\{w_1, w_2, w_3, w_4, w_5, w_6, w_7, w_8, w_9\}$.

We already have the uni-grams for this dictionary which are the probabilities of occurrence for each word (pattern). The joint probabilities are then calculated for every pair of words. The following formula is used to compute bi-grams:

$$Pr\,(w_i|\,w_j) = \frac{count(w_j, w_i)}{count(w_j)}\ i, j = 1, 2, \ldots, 9$$

2.3. Distance between Pitch Pattern Models

The Kullback Leibler (KL) divergence or relative entropy [18] is a non-commutative measure of similarity/dissimilarity between distributions or statistical models. If P and Q are two discrete probability distributions, the KL divergence of Q from P is:

$$D_{KL}\,(P\|\,Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}$$

In the previous subsection, we modeled pitch patterns for each dialect as a 2D discrete distribution. The next step is to compare dialect models to come up with a dialect distance measure. We used KL divergence for comparing the distributions which in this case has a closed form. The distance of dialect2 (D_2) from

dialect1 (D_1) is:

$$d(D_1, D_2) = \sum_{i=1}^{3} \sum_{j=1}^{3} P_1(i,j) \log \frac{P_1(i,j)}{P_2(i,j)}$$

where $P_1(i,j)$ and $P_2(i,j)$ are discrete distributions of dialect1 and dialect2, respectively. Note that $d(D_1, D_2)$ is not necessarily equal to $d(D_2, D_1)$. Therefore, we average the two distances to obtain separation assessment between two dialects.

3. Experimental Results and Evaluation

3.1. Distance Measures for Three Arabic Dialects

In this section, the use of distance assessment scheme is investigated for a corpus of three Arabic dialects: AE (United Arab Emirates), EG (Egypt), and SY (Syria). Our focus here is to keep the training data balanced, i.e., for each dialect almost 5 hours of train data from 32 male speakers is used. The computed distances using all the available data are as follows: d(AE,EG) = 0.0036, d(AE,SY)=0.0043, d(EG,SY) = 0.00018. The distances show that AE and SY have th widest separation, while EG and SY are the closest dialects. This is the same observation that resulted from previously proposed log-likelihood distances [13].

3.2. Evaluation

In order to show that the calculated distances are repeatable, we run the system on a subset of data which changes in a loop. In every iteration, different data collected from all the speakers is used. The results of 10 times running the distance measure algorithm using each time 1/10th of train data are summarized in Table 1. The first row of the table shows the distances when the whole data is used. In the second row mean and variance of the distances obtained from 10 experiments with subsets of data is shown.

Set	d(AE,SY) ($\times 10^{-3}$)	d(AE,EG) ($\times 10^{-3}$)	d(EG,SY) ($\times 10^{-3}$)
Whole Set	4.3	3.6	0.18
10 Subsets	4.4 ($\sigma = 0.6$)	3.7 ($\sigma = 0.5$)	0.22 ($\sigma = 0.07$)

Table 1: Mean and variances of 10 distance measures using subsets of the whole data set

In [13] we evaluated our log-likelihood score distribution distances with the results of an open-set GMM-based dialect classification task. 600 Mixtures and 26-dimensional MFCC features were used for classification. A confusion score was defined between each two dialects D_1 and D_2 as the sum of percentage of D_1 classified as D_2 and vice versa. Here we show that the pitch pattern based distance measures are consistent with log-likelihood measures. We also define a new separation assessment as the inverse of confusion score which obviously corresponds to the performance of dialect ID system. Fig. 3 shows the normalized dialect distances for the three Arabic dialects. The distances from comparing 3D pitch vector patterns can be compared to the log-likelihood score distribution distances. Joint PDF and bi-gram distances are also shown in the figure. These two correspond to 4D pitch vector pattern models. Inverse confusion distance can be used as a reference to show how well distance measures can predict dialect classification system's performance.

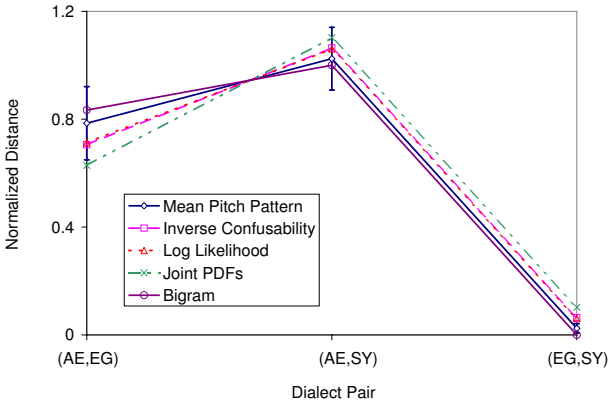

Figure 3: *Comparison between different dialect distance measures.*

3.3. Perceptive Evaluation

The correlation of presented objective dialect distance with human perception is shown in this subsection. For this experiment, two Egyptian subjects are used. Each subjective test consists of 30 sessions. In each session three 15 sec. conversations are presented from three different dialects (AE, EG, SY). One of the audio files is indicated as the reference. Listeners were asked to compare the two other utterances to the reference and on a scale of 1 (similar to the reference) to 10 (completely different from the reference) give two perceptual distances for each session. The reference dialect changes between sessions in a random way. To make the decisions as speaker independent as possible, 30 different speakers are used for all the sessions. The perceived distances between each two dialects are averaged across sessions to obtain one perceptive distance per listener. Since the native dialect of the subjects is Egyptian their judgment on comparing the other two dialects with their native dialect is more reliable. The resultant subjective distances from both listeners show that perceptually SY is closer to EG than AE to EG. This is the same result that we obtained from the proposed objective distance measures.

4. Conclusions

In this study, a method of assessing dialect separation based on comparing pitch movement patterns was proposed. 2D pitch slope vectors were first extracted from all the available train data for each dialect. The extracted vectors were later categorized into 9 patterns of pitch change. The probability of occurrence for each pattern were calculated to build statistical models. The obtained models were then compared using KL divergence. The proposed distance measure was evaluated for three Arabic dialects.The results show that AE dialect's pitch movements are completely distinguishable from the other two dialects (EG and SY). However EG and SY are more confiscable. Dialect Classification system's performance for these three dialects confirms the results of presented distance measure. The correlation of the distances with human perception was also investigated in a listener test. The proposed method of measuring dialect distance

has applications in dialect classification performance prediction as well as dialect data purity assessment.

5. References

[1] A. Curzan and M. Adams, *How English Works : A Linguistic Introduction*. Pearson Education Inc., 2006.

[2] P. A. Torres-Carrasquillo, T. P. Gleason, and D. A. Reynolds, "Dialect identification using Gaussian mixture models," in *ODYSSEY: The Speaker and Language Recognition Workshop*, Toledo, Spain, 2004, pp. 2977–300.

[3] R. Haung and J. H. L. Hansen, "Unsupervised discriminative training with application to dialect classification," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2444–2453, Nov. 2007.

[4] R. M. Gray, A. Buzo, A. H. G. JR., and Y. Matsuyama, "Distortion measures for speech processing," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 28, no. 4, pp. 367–376, Aug. 1980.

[5] D. Mansour and B. H. Juang, "A family of distortion measures based upon projection operation for robust speech recognition," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 37, no. 11, pp. 1659–1671, Nov. 1989.

[6] B. A. Carlson and M. A. Clements, "A projection-based likelihood measure for speech recognition in noise," *IEEE Trans. on Speech and Audio Processing*, vol. 2, no. 1, pp. 97–102, Jan. 1994.

[7] M. Padmanabhan, L. R. Bahl, D. Nahamoo, and M. A. Picheny, "Speaker clustering and transformation for speaker adaptation in speech recognition systems," *IEEE Trans. on Speech and Audio Processing*, vol. 6, no. 1, pp. 71–77, Jan. 1998.

[8] Y. Gao, M. Padmanabhn, and M. Picheny, "Speaker adaptation based on pre-clustering training speakers," in *Eurospeech*, 1997, pp. 2091–2094.

[9] S. R. Quackenbush, T. P. B. III, and M. A. Clements, *Objective Measures of Speech Quality*. Prentice-Hall, Inc., 1988.

[10] J. Wouters and M. W. Macon, "A perceptual evaluation of distance measures for concatenative speech synthesis," in *Proc. ICSLP*, Sydney, Australia, 1998.

[11] J. Kohler, "Multi-lingual phoneme recognition exploiting acoustic-phonetic similarities of sounds," in *Proc. ICSLP*, vol. 4, Oct. 1996, pp. 2195–2198.

[12] W. Heeringa, P. Kleiweg, C. Gooskens, and J. Nerbonne, "Evaluation of string distance algorithms for dialectology," in *Proc. of Workshop on Linguistic Distances*, Jul. 2006, pp. 51–62.

[13] M. Mehrabani and J. H. L. Hansen, "Dialect separation assessment using log-likelihood score distribution," in *Proc. Interspeech*, Brisbane, Australia, Sep. 2008.

[14] Y. K. Muthusamy, N. Jain, and R. A. Cole, "Perceptual benchmarks for automatic language identifications," in *Proc. ICASSP*, vol. 1, 1994.

[15] K. Kumpf and R. W. King, "Foreign speaker accent classification using phoneme-dependent accent discrimination models and comparisons with human perception benchmarks," in *Eurospeech*, 1997, pp. 2323–2326.

[16] R. Tong, B. Ma, D. Zhu, H. Li, and E. S. Chng, "Integrating acoustic, prosodic and phonotactic features for spoken language identification," in *Proc. ICASSP*, 2006, pp. 205–208.

[17] D. Talkin, *Speech Coding and Synthesis*. Amsterdam, Netherlands: Elsevier, 1995, ch. A Robust Algorithm for Pitch Tracking (RAPT). W.B. Kleijn and K.K. Paliwal (Eds.), pp. 495–518.

[18] S. Kullback, *Information Theory and Statistics*. New York: Dover Publications Inc., 1968.